Signatures of negative selection in the genetic architecture of human complex traits

Zeng, Jian; De Vlaming, Ronald; Wu, Yang; Robinson, Matthew R.; Lloyd-Jones, Luke R.; Yengo, Loic; Yap, Chloe X.; Xue, Angli; Sidorenko, Julia; McRae, Allan F.; Powell, Joseph E.; Montgomery, Grant W.; Metspalu, Andres; Esko, Tonu; Gibson, Greg; Wray, Naomi R.; Visscher, Peter M.; Yang, Jian

published in
Nature Genetics
2018

DOI (link to publisher)
10.1038/s41588-018-0101-4

document version
Publisher's PDF, also known as Version of record

ANALYSIS
https://doi.org/10.1038/s41588-018-0101-4
nature genetics

Signatures of negative selection in the genetic architecture of human complex traits

Jian Zeng[1], Ronald de Vlaming[2,3], Yang Wu[1], Matthew R. Robinson[1,4], Luke R. Lloyd-Jones[1], Loic Yengo[1], Chloe X. Yap[1], Angli Xue[1], Julia Sidorenko[1,5], Allan F. McRae[1], Joseph E. Powell[1], Grant W. Montgomery[1], Andres Metspalu[5], Tonu Esko[5], Greg Gibson[6], Naomi R. Wray[1,7], Peter M. Visscher[1,7] and Jian Yang[1,7]*

We develop a Bayesian mixed linear model that simultaneously estimates single-nucleotide polymorphism (SNP)-based heritability, polygenicity (proportion of SNPs with nonzero effects), and the relationship between SNP effect size and minor allele frequency for complex traits in conventionally unrelated individuals using genome-wide SNP data. We apply the method to 28 complex traits in the UK Biobank data ($N = 126{,}752$) and show that on average, 6% of SNPs have nonzero effects, which in total explain 22% of phenotypic variance. We detect significant ($P < 0.05/28$) signatures of natural selection in the genetic architecture of 23 traits, including reproductive, cardiovascular, and anthropometric traits, as well as educational attainment. The significant estimates of the relationship between effect size and minor allele frequency in complex traits are consistent with a model of negative (or purifying) selection, as confirmed by forward simulation. We conclude that negative selection acts pervasively on the genetic variants associated with human complex traits.

Dissecting the genetic architecture of complex traits is important for understanding the genetic basis of phenotypic variation and evolution. For a fitness-related complex trait, natural selection plays an important role in shaping its genetic architecture[1], which in turn provides information to infer the action of natural selection. Given that most traits are polygenic, natural selection is likely to act simultaneously on many trait-associated variants that have pleiotropic effects on fitness[2,3]. Unlike a selective sweep model[4], in which there are often a limited number of mutations under relatively strong selection, it is difficult to detect the selection signals in polygenic traits because the selection pressure is spread over many variants of small effect. However, evidence for natural selection can be inferred from the relationship between effect size and minor allele frequency (MAF) for all genome-wide variants. For example, mutations that are deleterious to fitness are selected against and thus kept at low frequencies by negative selection (also known as purifying selection), resulting in a negative relationship between effect size and MAF[5,6]. The estimation of the joint distribution of effect size and MAF can be used to detect signatures of natural selection[7] and thereby infer the relationship between a complex trait and fitness.

Genome-wide association studies (GWAS) have detected thousands of SNPs associated with complex traits, which have helped to characterize the genetic architecture of these traits[8]. However, the genome-wide significant SNPs discovered in GWAS jointly explain only a fraction of the heritability, because many SNPs with small effects are yet to be detected[9]. Furthermore, a proportion is missed due to the incomplete linkage disequilibrium (LD) between causal variants and SNP markers[9]. To address the 'missing heritability' problem[9,10] in GWAS, mixed linear model (MLM) approaches have been used to estimate the variance explained by all SNPs used in a GWAS. GREML is a prevailing class of MLM-based approaches, in which all SNP effects are fitted together as random effects[11]. GREML analyses using common SNPs (MAF > 1%) have uncovered a large proportion of the missing heritability for height[12], body mass index (BMI)[12], and psychiatric disorders[13]. The GREML method assumes that all SNPs have an effect on the trait[11] and thus does not allow us to estimate the degree of polygenicity (i.e., the proportion of SNPs with nonzero effects). Bayesian multiple regression is another class of MLM-based methods that enable us to make posterior inferences about polygenicity by assuming SNP effects are drawn from a mixture distribution of zero and nonzero components[14,15]. Bayesian MLM methods have been widely used in livestock and plant breeding[16] and have attracted increasing attention in humans for characterizing the genetic architecture of complex traits and diseases[15,17,18]. However, neither GREML nor Bayesian MLM explicitly models the relationship between effect size and MAF for complex traits.

In this study, we developed a Bayesian MLM method that can simultaneously estimate SNP-based heritability, polygenicity, and the joint distribution of effect size and MAF in conventionally unrelated individuals using GWAS data. We applied the method to 28 complex traits in the UK Biobank (UKB) data[19] and to 27,869 gene expression traits in the Consortium for the Architecture of Gene Expression (CAGE) data[20].

Results

Method overview. Under the Bayesian MLM framework, we modeled the relationship between effect size and MAF using the following mixture distribution as a prior for each SNP effect:

$$\beta_j \sim N\left(0, \left[2p_j\left(1-p_j\right)\right]^S\sigma_\beta^2\right)\pi + \phi\left(1-\pi\right)$$

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia. [2]School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [3]Erasmus University Rotterdam Institute for Behavior and Biology, Rotterdam, The Netherlands. [4]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [5]Estonian Genome Center, University of Tartu, Tartu, Estonia. [6]School of Biological Sciences and Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, GA, USA. [7]Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia. *e-mail: jian.yang@uq.edu.au

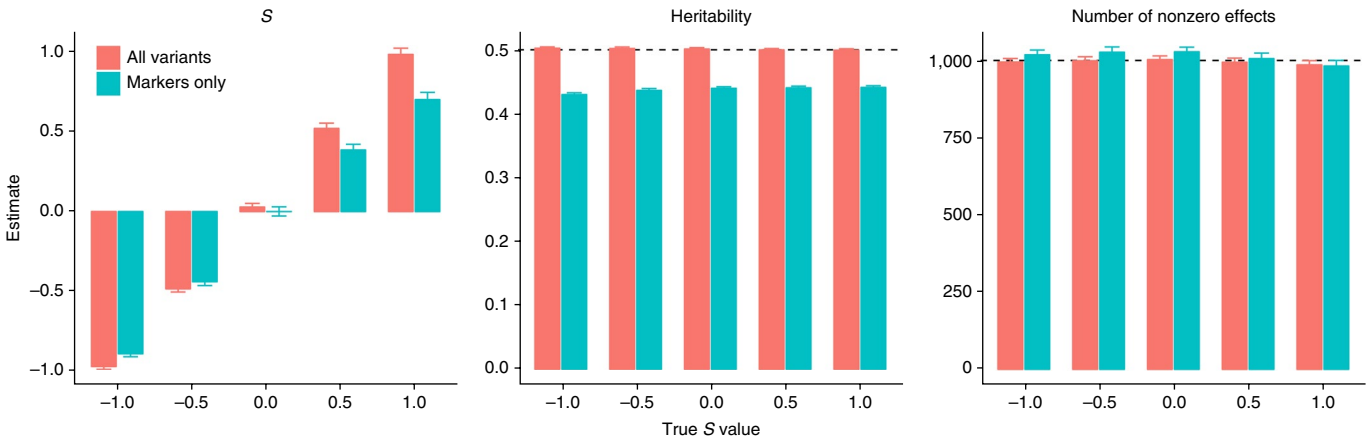

Fig. 1 | Estimation of the genetic architecture parameters for a simulated trait using the ARIC + GENEVA data. Results are the mean estimates with s.e.m. (error bars) over 100 simulation replicates for a spectrum of *S* parameter values. Color indicates the results of BayesS with both causal variants and SNP markers (red) or with SNP markers only (blue). The heritability parameter at the 1,000 randomly selected causal variants is 0.5. The number of nonzero effects ($m_{NZ} = m\pi$ with m being the total number of SNPs) is the number of SNPs with nonzero effects averaged over MCMC iterations.

where β_j is the allelic substitution effect of an SNP j, p_j is the MAF of the SNP, σ^2_β is a constant factor (i.e., variance of SNP effects under a neutral model), ϕ is a point mass at zero, and π is the proportion of SNPs with nonzero effects (i.e., polygenicity). The variance of β_j when β_j is nonzero, $\sigma^2_j = [2p_j(1-p_j)]^S\sigma^2_\beta$, is defined over all possible values of S, which relates the variance of SNP effects to a MAF that is equivalent to a previous class of MLMs when $S = 0$ and $S = -1$[21–23] (see part 1 of the Supplementary Note for a proof of model equivalence). When $S = 0$, the effect size is independent of MAF (neutral model). If there is selection, the effect size can be positively ($S > 0$) or negatively ($S < 0$) related to MAF. In this model (referred to as BayesS), we used a gradient-based sampling algorithm, Hamiltonian Monte Carlo[24], to sample S from the posterior distribution and used Gibbs sampling for other parameters in the model by assuming conjugate priors (Methods). Furthermore, we used a parallel computing strategy following Fernando et al.[25] to scale the analysis for very large sample sizes ($N > 100{,}000$). Details of the sampling scheme and the parallel computing strategy are given in parts 2 and 3 of the Supplementary Note. In the hypothesis test against $S = 0$, two approaches were used to control false positives (Methods and part 4 of the Supplementary Note), justified by simulation (Supplementary Figs. 1 and 2).

Assessing parameter estimation through simulations. We used simulations based on real GWAS genotype data from the Atherosclerosis Risk in Communities (ARIC) and GENEVA Diabetes study[12] from dbGaP to assess the robustness of our method in estimating the parameters $\boldsymbol{\theta} = [S, h^2_{SNP}, \pi]$, where π is assessed by the number of SNPs with nonzero effects (i.e., $m_{NZ} = \pi m$ with m being the total number of SNPs; see Methods). The ARIC + GENEVA data consisted of 12,942 unrelated individuals and 564,959 Affymetrix SNPs with MAF > 1% after quality controls. In our simulation, 1,000 SNPs were chosen at random to be causal variants, with their effects related to MAF through an S value ranging from −1 to +1 in different scenarios.

Results showed that when both causal variants and SNP markers were fitted in the analysis, $\hat{\boldsymbol{\theta}}$ from BayesS was unbiased with respect to the true parameters (Fig. 1). When the causal variants were not included in the analysis, both $\hat{h}^2_{SNP}$ and the absolute value of $\hat{S}$ were slightly underestimated, due to imperfect tagging, an issue similar to that discussed in Yang et al.[11]. For polygenicity, $\widehat{m}_{NZ}$ was an approximately unbiased estimate of the number of causal variants (m_C) because the common causal variants (MAF > 1%) were well tagged by the common SNPs. Regardless of whether the causal variants were observed or masked, the standard error (s.e.) for $\hat{S}$, $\hat{h}^2_{SNP}$, and $\widehat{m}_{NZ}$ were consistent with the standard deviation (s.d.) of the estimates from 100 simulation replicates (Supplementary Table 1). We also assessed the method under more extreme S values, as well as under different levels of heritability, numbers of causal variants, and sample sizes. The results were similar to those reported above, except that the m_{NZ} estimate tended to be larger than m_C when Nh^2/m_C was small, where N is the sample size and h^2 is the trait heritability (Supplementary Fig. 3), likely due to the lack of power to distinguish between fitting a single causal variant and fitting multiple SNPs to capture the genetic variance in a genomic region. Thus, in practice, $\widehat{\pi}$ needs to be estimated in a large sample (e.g., N values in the tens of thousands or larger) and interpreted as the proportion of non-null SNPs. The extent to which it reflects the proportion of causal variants depends on how well the causal variants are tagged by the SNPs and the power. Nevertheless, the overestimation of m_C did not inflate the estimate of S (Supplementary Fig. 3).

We also performed simulations based on whole-genome sequencing (WGS) data from the UK10K project[26] (Methods) and showed that the estimates were unbiased in the presence of rare variants and that the estimates from downsampled array SNPs were shrunk towards zero due to the imperfect tagging of causal variants by array SNPs (Supplementary Fig. 4). When the S parameter increased from negative to positive, the proportion of genetic variance due to common causal variants also increased, resulting in increased tagging of causal genetic variation by array SNPs (Supplementary Fig. 4). Moreover, the method appeared to be robust to heterogeneity of LD in the genome (part 5 of the Supplementary Note and Supplementary Fig. 4).

Analysis of 28 complex traits in the UK Biobank data. We applied the BayesS method to 36 complex traits on 126,545 unrelated individuals of European ancestry in the UKB[19] with 483,634 Affymetrix SNPs (MAF > 1%) after quality controls (Methods). Of the 36 traits, 21 had $N > 100{,}000$. Two commonly used long-chain diagnostic tests were adopted to assess the convergence of the Markov chain Monte Carlo (MCMC) algorithm (part 6 of the Supplementary Note). Traits that did not pass our convergence tests were those with the smallest sample sizes, $\hat{h}^2_{SNP}$ close to zero, or both (Supplementary Fig. 5). We focus on the results of 28 traits that passed both convergence tests for all of the three genetic architecture parameters (Supplementary Fig. 6).

Comparison of genetic architectures between height and BMI. The genetic architectures of height and BMI have been relatively well studied compared to other complex traits[23,27–31]. Thus, it is revealing to compare our results for height and BMI (Fig. 2) with the previous findings. Both traits have a large sample size in the UKB: $N = 126{,}545$ for height and $N = 126{,}389$ for BMI. A negative S was detected with extremely high significance for both traits ($P = 1.5 \times 10^{-106}$ for height and $P = 3.2 \times 10^{-13}$ for BMI), meaning that lower-MAF variants tend to have larger effect sizes. Because our method models the relationship between MAF and variance of SNP effects, it cannot distinguish whether the negative estimate of S is due to an enrichment of trait-increasing or -decreasing alleles with larger effects at lower frequencies. Nevertheless, the estimated negative relationship between MAF and variance of SNP effects suggests that both height- and BMI-associated SNPs have been under negative selection, in line with the conclusions from recent studies[23,30] (see part 7 of the Supplementary Note for more implications of the results).

Inference on natural selection. Of the 28 traits that passed our convergence tests, 23 traits (including reproductive, cardiovascular, and anthropometric traits, as well as educational attainment) had significant negative S estimates with posterior probability $\Pr(S < 0 | \text{ data}) = 1$ and $P < 0.05/28$ (Supplementary Table 2), providing strong evidence that the genetic variants associated with these traits have been under selection. The estimates of S over traits ranged from −0.609 (age at menopause) to 0.012 (fluid intelligence score) with mean −0.355, median −0.365, and s.d. 0.109. Notably, all the significant estimates of S were negative (see below for forward simulation to infer the type of selection from the sign of S). The magnitude of $\hat{S}$, i.e., $|\hat{S}|$, reflects the strength of selection on the trait-associated SNPs. Traits with the largest $|\hat{S}|$ are related to fertility and heart function (Fig. 3), including age at menopause ($\hat{S} = -0.609$, s.e. = 0.073), pulse rate ($\hat{S} = -0.486$, s.e. = 0.048), waist circumference adjusted for BMI ($\hat{S} = -0.426$, s.e. = 0.036), and waist–hip ratio adjusted for BMI ($\hat{S} = -0.419$, s.e. = 0.048). It has been reported that waist circumference and waist–hip ratio are associated with cardiovascular events[32], and the latter is strongly correlated with pregnancy rate[33]. Age at menopause has a substantial impact on lifetime fertility and a recent study has found molecular links between the onset and end of reproductive lifespan[34]. Other reproductive and cardiovascular traits, such as age at menarche, age at first live birth, and blood pressure, had relatively high $|\hat{S}|$ values as well. Thus, our results suggest that reproductive and cardiovascular traits are closely related to fitness and that the SNPs associated with these traits have been under relatively stronger selection than SNPs associated with other traits.

Height ($\hat{S} = -0.422$), handgrip strength (right: −0.423, left: −0.380), lung function–related traits ($\hat{S}$ values ranging from −0.401 to −0.328), heel bone mineral density (−0.381), and basal metabolic rate (−0.359) had moderate to high $|\hat{S}|$ values (Fig. 3 and Supplementary Table 2). Two diseases or disorders, type 2 diabetes (T2D) and depression (DEP), had negative $\hat{S}$ values but the P values did not reach significance (0.05/28), although the posterior probability of $S < 0$ for T2D was as high as 0.983. However, the power to detect a significant $\hat{S}$ may be lower for these disease traits, given that the number of cases was less than 10,000 for each. A recent GWAS meta-analysis of ~110,000 individuals with SNP data imputed from a WGS sample of ~2,700 individuals did not find evidence of natural selection on T2D-associated variants[35], which may also be due to insufficient power. Fluid intelligence (FI) score was the only trait with $\hat{S}$ at almost zero ($\hat{S} = 0.012$, s.e. = 0.096); however, there was strong evidence of negative selection on the SNPs associated with educational attainment ($\hat{S} = -0.335$, s.e. = 0.055), which is thought to be a proxy of intelligence. Given that the genetic correlation between educational attainment and FI was as high as 0.665 (s.e. = 0.052), estimated from a bivariate LD score regression[36], the fact that we did not detect a signal of selection for FI-associated SNPs may be due to the limited statistical power. It is also possible that both positive and negative selection pressures have acted on FI-associated SNPs, resulting in a flat relationship between MAF and effect size (see below for more discussion).

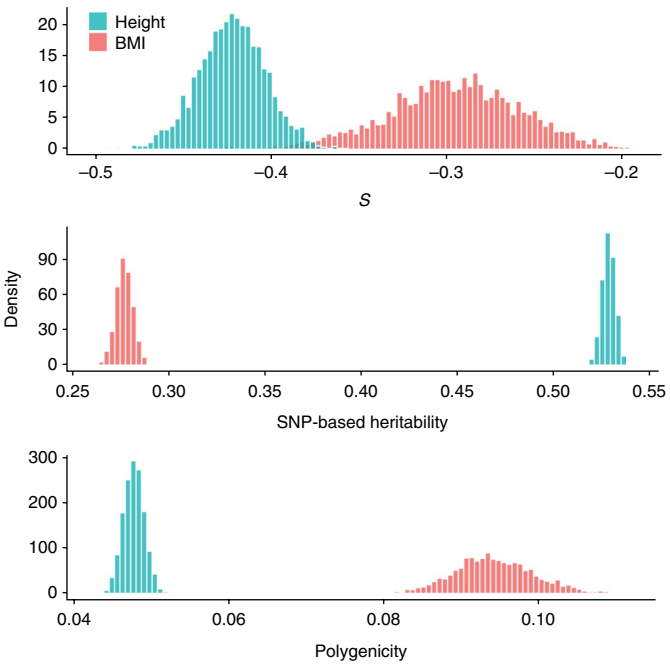

Fig. 2 | Posterior distributions of the genetic architecture parameters for height versus BMI using data from UKB. S measures the relationship between effect size and MAF. Polygenicity is defined as the proportion of SNPs with nonzero effects. The sample size is 126,545 for height and 126,389 for BMI.

For traits with a significant estimate of S, we demonstrated the relationship between effect size and MAF by a plot of the cumulative proportion of genetic variance explained by SNPs (cGVE) against MAF (Fig. 4), where MCMC samples of SNP effects were used to compute cGVE for SNPs with MAF smaller than a threshold on the x axis (part 8 of the Supplementary Note). Under an evolutionarily neutral model and assuming a constant population size, cGVE is linearly proportional to MAF[37,38] (i.e., the integral of per-SNP additive genetic variance over a MAF bin is a constant), and therefore the area under the curve (AUC) is expected to be 0.5. All traits with significant estimates of S had the curve of cGVE above the diagonal line, with $|\hat{S}|$ highly correlated with the AUC (Pearson's correlation = 0.896). Demographic events such as population bottlenecks and rapid expansions could lead to changes in the distribution of cGVE[39,40] but would not bias the estimate of S as demonstrated by a coalescent simulation[41] (Methods and Supplementary Fig. 7). Nevertheless, our observation of high correlation between $|\hat{S}|$ and the AUC (Fig. 4) implies that the impact of demographic events on the distribution of cGVE is likely to be modest.

Inference on SNP-based heritability. The 28 traits had low to moderate estimates of $\hat{h}^2_{SNP}$, with mean 22.4%, median 21.5%, and s.d. 9.5% (Supplementary Table 3). Note that traits with $\hat{h}^2_{SNP}$ close to zero had failed in MCMC convergence tests (Supplementary Fig. 5), and therefore the mean $\hat{h}^2_{SNP}$ estimate across traits is likely to be inflated. Besides height ($\hat{h}^2_{SNP} = 52.7\%$), traits with the highest $\hat{h}^2_{SNP}$ (Supplementary Fig. 8) included basal metabolic rate (33.8%), which has been reported to lie between 0.2 and 0.4 in model animals[42], and male pattern baldness (33.7%), which has been

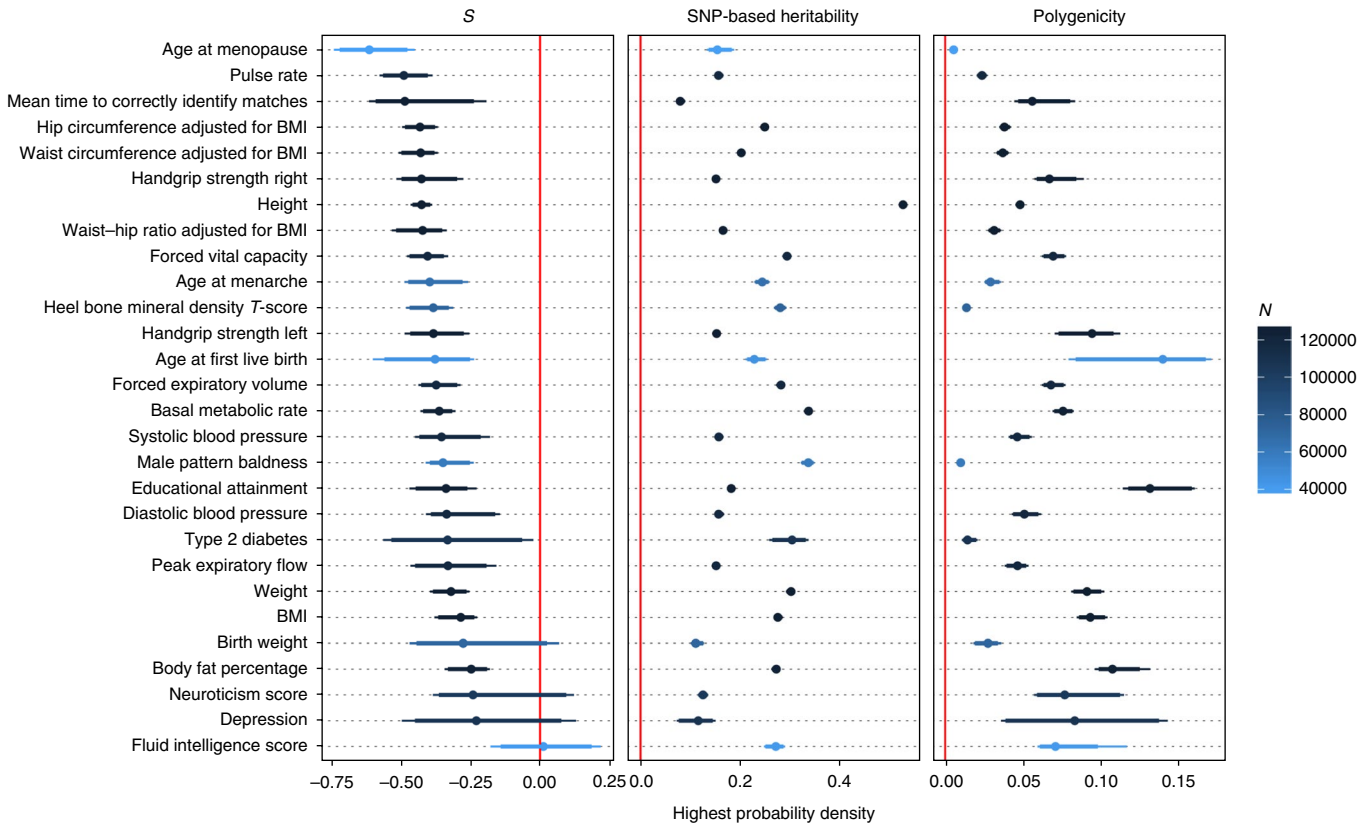

Fig. 3 | BayesS estimates of the genetic architecture parameters for the UKB traits. Results are for the 28 UKB complex traits that passed convergence tests on the MCMC chain. Points indicate the posterior modes estimated from the MCMC sample; bold lines represent 90% credible intervals (highest posterior density), and thin lines represent 95% credible intervals. Sample size *N* for each trait is shown by the color gradient. Polygenicity is defined as the proportion of genome-wide SNPs with nonzero effects on the trait.

reported to be a highly heritable trait in twin studies[43]. Traits with the lowest $\hat{h}^2_{SNP}$ included mean time to correctly identify matches (8.0%), birth weight (11.1%), DEP (11.6%), and neuroticism score (12.5%), in line with the low estimates of $\hat{h}^2_{SNP}$ from previous studies of DEP[44] and neuroticism score[45]. Given that most published estimates were obtained using whole-genome imputed SNPs, they are likely to be slightly higher than our estimates based on array SNPs. For example, a recent study[46] on educational attainment in UKB gave an estimate of 21% (s.e.=0.6%), slightly higher than our estimate of 18.2% (s.e.=0.4%). Our estimate of 52.7% (s.e.=0.3%) for height was slightly but not substantially lower than that of 56% (s.e.=2.3%) in Yang et al.[23]. For BMI, our estimate of 27.6% (s.e.=0.4%) was highly consistent with that of 27% (s.e.=2.5%) in Yang et al.[23]. Across traits, $\hat{h}^2_{SNP}$ seemed to be independent of either $\hat{S}$ or $\hat{\pi}$, but the s.e. values of $\hat{S}$ and $\hat{\pi}$ decreased as $\hat{h}^2_{SNP}$ increased (Supplementary Fig. 9).

Inference on polygenicity. The distribution of $\hat{\pi}$ had mean 5.9%, median 5.4%, and s.d. 3.5% across traits, and ranged from 0.6% (s.e.=0.1%) to 14.0% (s.e.=1.3%; Supplementary Table 4). This suggests that all 28 complex traits are polygenic, with ~30,000 common SNPs with nonzero effects, on average. Notably, age at menopause, the trait with highest magnitude of $\hat{S}$ (−0.609), had the lowest estimate of polygenicity ($\hat{\pi}$=0.6%, s.e.=0.1%; Supplementary Fig. 8). Age at first live birth had the highest $\hat{\pi}$ (14.0%) but with a relatively large s.e. (2.5%). It was followed by educational attainment ($\hat{\pi}$=13.2%, s.e.=1.3%), which is reasonable because it is likely to be made up of several subphenotypes so that many SNPs have an effect. We examined the robustness of results above by repeating the UKB analyses with smaller sample sizes (Supplementary Fig. 10) and by simulations with the estimated genetic architectures from the UKB traits (Supplementary Table 5, part 9 of the Supplementary Note, and Supplementary Figs. 10–12).

The signs of *S* under different types of natural selection. Besides detecting selection and quantifying its strength on the trait-associated SNPs, the sign of *S* allows us to further infer the type of selection. To demonstrate this, we used forward simulations to simulate common types of natural selection for a quantitative trait by relating the normally distributed phenotype to fitness through a hypothetical function (Methods and part 10 of the Supplementary Note). In this setting, the trait itself was under selection, and the sign of *S* distinguished stabilizing selection ($S<0$) from directional and disruptive selection ($S>0$) on the trait (Fig. 5 and Supplementary Fig. 13); this property held under a demographic model[47] (part 11 of the Supplementary Note and Supplementary Fig. 14). Our result also implied that when the trait-associated variants have pleiotropic effects on fitness, the sign of *S* distinguishes negative selection ($S<0$) from positive selection ($S>0$) on the variants (Fig. 5), confirmed by simulations with variants of direct deleterious or beneficial effects on fitness (Supplementary Fig. 15).

Analysis of gene expression traits. To demonstrate that our method can be applied to less polygenic traits, we analyzed gene expression data from CAGE[20], which consisted of 36,778 gene expression probes in peripheral blood from 1,748 unrelated individuals of European ancestry and 1,066,738 HapMap3[48] SNPs with MAF>1% (Methods). To facilitate the analysis of all 36,778 probes, we developed a nested version of the BayesS method (Methods

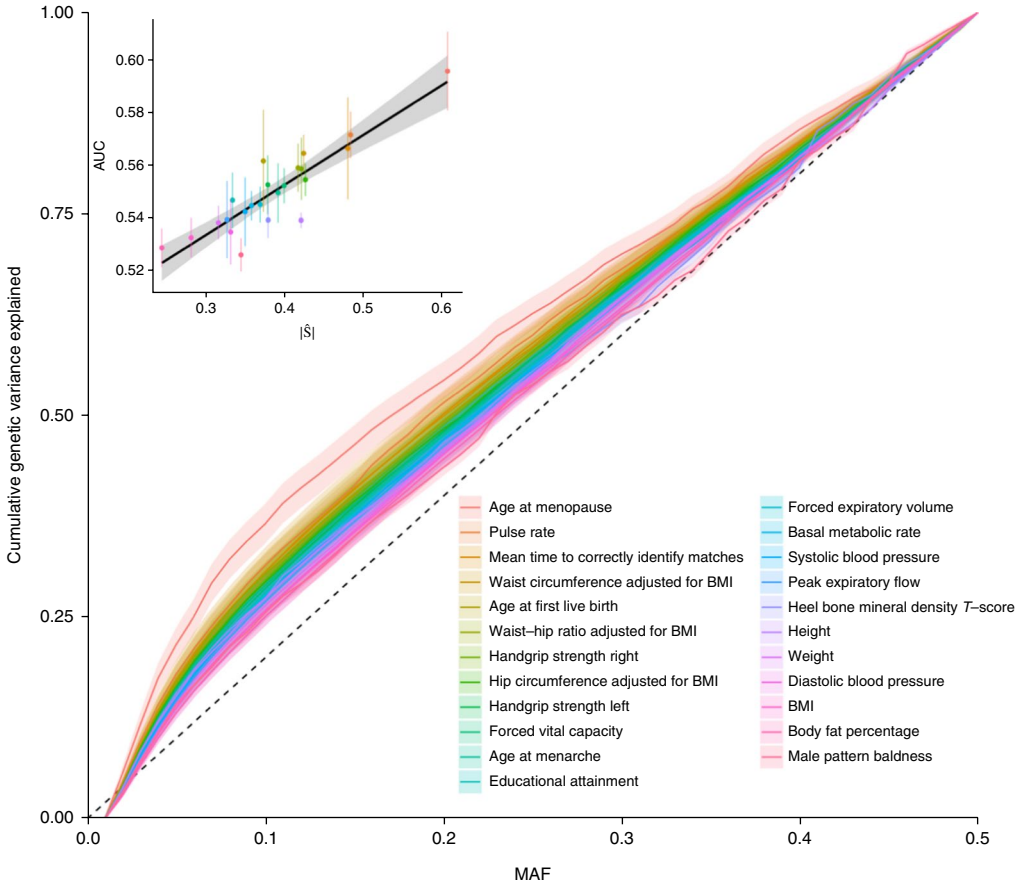

Fig. 4 | cGVE by SNPs with MAF smaller than a threshold on the *x* axis. Results are for the 23 UKB complex traits for which the estimates of *S* are significantly different from zero. Data are shown along the x axis for MAFs starting at 0.01, and move toward 0.05 with a step size of 0.01. The posterior mean of the cGVE is computed from the SNPs with MAF ≤ *x* in each step and two adjacent steps are connected by a straight line. The shaded area shows the posterior standard error. Inset: the relationship between the posterior mean of the AUC of the cGVE (posterior s.e. indicated by the error bar) and $|\hat{S}|$ across traits (Pearson's correlation $r = 0.896$). The dark line is the regression line and shaded area shows the 95% confidence interval. $N = 38{,}642$ to 126,545 biologically independent samples.

and parts 12–15 of the Supplementary Note) and evaluated its performance using both real and simulated data (Supplementary Figs. 16–18). We found that the estimates of h^2_{SNP} and π for most probes were lower than those for UKB traits (Supplementary Figs. 18–20). Applying a Bonferroni correction for the number of probes mapped to the genome, we observed significant $\hat{S}$ values for 32 probes ($P < 0.05/21{,}303 = 2.3 \times 10^{-6}$; Supplementary Fig. 21 and Supplementary Table 6), all of which had a negative $\hat{S}$ value (mean = –1.26, s.d. = 0.19), mapped to 30 unique genes (Supplementary Fig. 22). Of these 32 probes, 29 had expression quantitative trait loci (eQTL) at $P < 5 \times 10^{-8}$, and $\hat{S}$ for all 29 probes became nonsignificant after removing SNPs within 2 Mb around each probe (Supplementary Table 7), consistent with the result from forward simulation for traits with genetic architecture of 'a large cis-eQTL + polygenic effects' (part 16 of the Supplementary Note and Supplementary Fig. 23). The forward simulation also showed that for gene expression traits with a large cis-eQTL, the *S* parameter for individual traits varied substantially under a neutral model, but the mean of *S* was still informative to infer the type of selection across traits (Supplementary Fig. 23). Thus, our observation that $\hat{S}$ was negative for all 32 detected probes is consistent with a model of negative selection on the genetic variants associated with gene expression traits. In addition, there was little correlation between selection on the regulatory and coding regions of a gene (Supplementary Fig. 24), consistent with the findings in Torgerson et al.[49].

Discussion

We introduced a method (called BayesS) to infer the action of natural selection on the genetic variants underlying a complex trait. By estimating the relationship between the variance of SNP effects and MAF (i.e., the parameter *S*) using genome-wide SNPs, we detected significant signatures of natural selection ($S \neq 0$) for 23 of 28 complex traits in the UKB data, with the strongest selection signals from SNPs associated with reproductive and cardiovascular traits, followed by those associated with height, handgrip strength, lung function, and other anthropometric traits, as well as educational attainment (Fig. 3). Together with the high prevalence of selection signals across traits (23/28 = 82%), our observation of high degree of polygenicity (~6% on average) underlines the role of pleiotropy in the action of natural selection.

Our observation that all the significant estimates of *S* were negative (Fig. 3) is consistent with a model of negative selection (Fig. 5). Evidence of negative selection has been reported for height and BMI previously, from estimating the variance explained by variants stratified by MAF and LD scores (the GREML–LDMS approach)[23]. When we applied the GREML–LDMS approach to the UKB data imputed from Haplotype Reference Consortium[50], we also found that variance explained by rare SNPs (MAF < 1%) was larger than expected under a neutral model for height, BMI, waist-to-hip ratio, and diastolic blood pressure (Supplementary Table 8). The GREML–LDMS analysis, however, cannot distinguish whether the excess of variance explained by rare variants is a result of negative

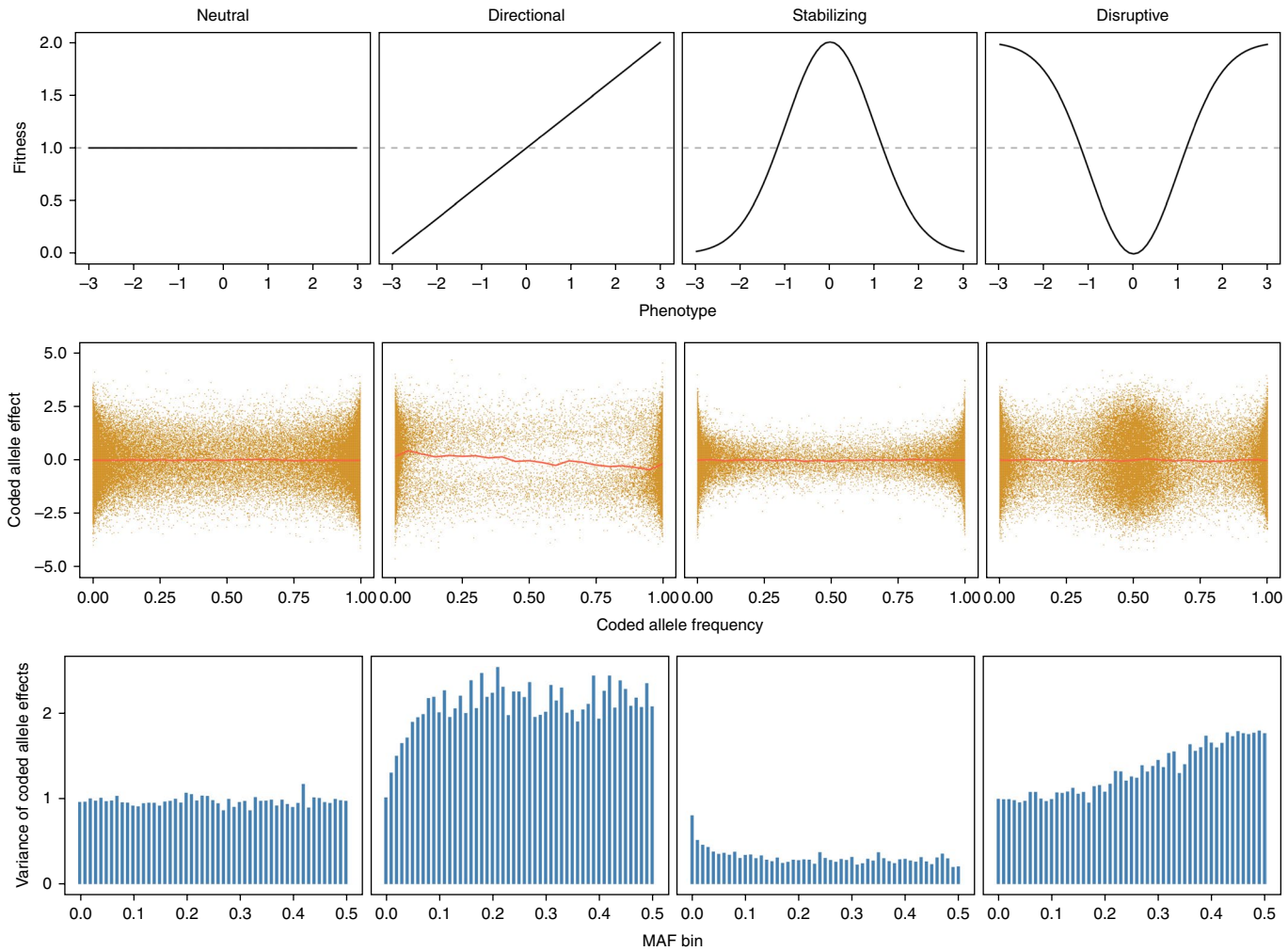

Fig. 5 | Forward simulations with different types of selection. A quantitative trait was generated based on a simulated chromosome segment of 10 Mb (5% causal and 95% neutral mutations in each generation). The trait heritability is 0.1. Top: the functions used to relate the phenotype (normally distributed) to fitness in different modes of selection: neutral variation, directional selection, stabilizing selection, and disruptive selection. Middle: the joint distributions of the coded allele effects and frequencies, where the coded allele at each causal variant was chosen at random from the derived and ancestral alleles, and the red line shows the means of coded allele effects in allele frequency intervals of 0.05. Bottom: relationships between the variance of the coded allele effects and MAF. Results are collected from 200 replicates of simulation.

selection or an excess of rare variants due to rapid population expansion[39,40]. Our BayesS approach, which models the variance of SNP effects conditional on MAF, is therefore expected to be robust to the abundance of rare variants, as confirmed by the coalescent simulation (Methods and Supplementary Fig. 7) and by an additional forward simulation under a commonly used demographic model[47] (Supplementary Fig. 14).

There is an increasing body of literature supporting the hypothesis of widespread signatures of polygenic selection on genetic variants associated with complex traits[3,29,51–54]. The metrics used in these studies, such as F_{ST}[52] and singleton density score[54], are specialized to detect signals of positive selection, whereas our method uses the S parameter to infer the type of selection that has played a predominant role in shaping the relationship between MAF and effect size. Our conclusions of widespread negative selection are consistent with findings of a recent study that used an extended LD score regression approach to estimate the effects of LD-based annotations on per-SNP heritability[55]. It is likely that both positive and negative selection have acted on the genetic variants associated with a complex trait and that the former is weaker and/or less frequent than the latter so that the population evolves (slowly) in one direction by directional selection with a persistent constraint on extreme phenotypes by stabilizing selection. This hypothesis is supported by the results from an additional forward simulation in the presence of both positive and negative selection (part 17 of the Supplementary Note), which showed that S increased from negative to positive when the strength of positive selection increased relative to that of negative selection (Supplementary Fig. 25).

Because we only included the array SNPs in the UKB analysis, our inference on natural selection is likely to be conservative due to the imperfect tagging of causal variants by array SNPs, as demonstrated by the simulation based on the UK10K WGS data (Supplementary Fig. 4). The median estimate of S of all the UKB traits was −0.37 (s.d. = 0.11) with a minimum estimate of −0.609 (s.e. = 0.073) for age at menopause, which is not as low as the default value used in the GREML method implemented in the GCTA software (i.e., $S = -1$)[11,56]. These observations may be because most of the UKB traits are not strongly correlated with fitness; the strength of stabilizing selection on fitness has been limited in human populations; a substantial proportion of causal variation has not been captured by SNPs; and/or the signatures of negative and positive selection are cancelled out by each other (see the discussion above).

Given that most complex traits have negative estimates of the relationship between variance of SNP effects and MAF, future GWAS based on WGS or imputed data with large sample sizes are expected to discover an increasing number of rare variants with large effects.

We conclude with several caveats. First, the polygenicity estimate ($\hat{\pi}$) only approximately reflects the actual fraction of causal variants, unless the sequence variants are observed and the sample size is sufficiently large. Nevertheless, $\hat{\pi}$ can be used to compare the levels of polygenicity across traits if the proportion of variation at causal variants tagged by array SNPs and the S parameter are not too different between traits. Second, the power of detecting a signal of natural selection (i.e., testing against $S = 0$) may improve if WGS or imputed sequence data that include a large number of rare variants are used in the analysis. However, further computational optimization is needed to run BayesS on all the WGS variants in a large cohort like the UKB. Finally, it is important to note that our study detected signatures of natural selection on genetic variants associated with complex traits. Whether the traits themselves have been under selection is not known.

URLs. GCTA, http://cnsgenomics.com/software/gcta. GCTB, http://cnsgenomics.com/software/gctb. SLiM, https://messerlab.org/slim. msprime, https://github.com/jeromekelleher/msprime. PLINK, https://www.cog-genomics.org/plink2. ldsc, https://github.com/bulik/ldsc. UK Biobank, http://www.ukbiobank.ac.uk.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0101-4.

Received: 31 May 2017; Accepted: 5 March 2018;
Published online: 16 April 2018

References

1. Johnson, T. & Barton, N. Theoretical models of selection and mutation on quantitative traits. *Phil. Trans. R. Soc. Lond. B* **360**, 1411–1425 (2005).
2. Hancock, A. M. et al. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc. Natl. Acad. Sci. USA* **107**(Suppl 2), 8924–8930 (2010).
3. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208–R215 (2010).
4. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
5. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
6. Eyre-Walker, A. Evolution in health and medicine Sackler colloquium: genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci. USA* **107**(Suppl 1), 1752–1756 (2010).
7. Mancuso, N. et al. The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* **48**, 30–35 (2016).
8. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
9. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
10. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
11. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
12. Yang, J. et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
13. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* **17**, 782–790 (2014).
14. Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**, 186 (2011).
15. Moser, G. et al. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* **11**, e1004969 (2015).
16. de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. & Calus, M. P. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 327–345 (2013).
17. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
18. Lloyd-Jones, L. R. et al. Inference on the genetic basis of eye and skin color in an admixed population via Bayesian linear mixed models. *Genetics* **206**, 1113–1126 (2017).
19. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
20. Lloyd-Jones, L. R. et al. The genetic architecture of gene expression in peripheral blood. *Am. J. Hum. Genet.* **100**, 228–237 (2017).
21. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
22. Lee, S. H. et al. Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151–1155 (2013).
23. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
24. Neal, R. M. MCMC using Hamiltonian dynamics. in *Handbook of Markov Chain Monte Carlo* (eds Brooks, S., Gelman, A., Jones, G. & Meng, X.-L.) 113–162 (CRC Press, Boca Raton, FL, 2011).
25. Fernando, R. L., Dekkers, J. C. & Garrick, D. J. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet. Sel. Evol.* **46**, 50 (2014).
26. Walter, K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
27. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
28. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
29. Robinson, M. R. et al. Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.* **47**, 1357–1362 (2015).
30. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
31. Turcot, V. et al. Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat. Genet.* **50**, 26–41 (2018).
32. de Koning, L., Merchant, A. T., Pogue, J. & Anand, S. S. Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: meta-regression analysis of prospective studies. *Eur. Heart J.* **28**, 850–856 (2007).
33. Wass, P., Waldenström, U., Rössner, S. & Hellberg, D. An android body fat distribution in females impairs the pregnancy rate of in-vitro fertilization-embryo transfer. *Hum. Reprod.* **12**, 2057–2060 (1997).
34. Day, F. R. et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.* **47**, 1294–1303 (2015).
35. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
36. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
37. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
38. Visscher, P. M., Goddard, M. E., Derks, E. M. & Wray, N. R. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol. Psychiatry* **17**, 474–485 (2012).
39. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
40. Uricchio, L. H., Zaitlen, N. A., Ye, C. J., Witte, J. S. & Hernandez, R. D. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Res.* **26**, 863–873 (2016).
41. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
42. Konarzewski, M. & Książek, A. Determinants of intra-specific variation in basal metabolic rate. *J. Comp. Physiol. B* **183**, 27–41 (2013).
43. Nyholt, D. R., Gillespie, N. A., Heath, A. C. & Martin, N. G. Genetic basis of male pattern baldness. *J. Invest. Dermatol.* **121**, 1561–1564 (2003).
44. Hyde, C. L. et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48**, 1031–1036 (2016).
45. de Moor, M. H. et al. Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry* **72**, 642–650 (2015).

46. Davies, G. et al. Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N = 112 151). *Mol. Psychiatry* **21**, 758–767 (2016).
47. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA* **108**, 11983–11988 (2011).
48. Altshuler, D. M. et al. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
49. Torgerson, D. G. et al. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet.* **5**, e1000592 (2009).
50. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
51. Hernandez, R. D. et al. Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920–924 (2011).
52. Turchin, M. C. et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44**, 1015–1019 (2012).
53. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).
54. Field, Y. et al. Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
55. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
56. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

Acknowledgements

We thank The University of Queensland's Research Computing Centre (RCC) for its support in this research. We thank F. Zhang for building the website for the software tool GCTB. This research was supported by the Australian Research Council (DP160101343, DP160101056, DP160103860, and DP160102400), the Australian National Health and Medical Research Council (1107258, 1078901, 1078037, 1083656, 1078399, 1046880, and 1113400), the US National Institutes of Health (MH100141, GM099568, ES025052, and AG042568), and the Sylvia & Charles Viertel Charitable Foundation (Senior Medical Research Fellowship). R.d.V. acknowledges funding from an ERC consolidator grant (647648 EdGe, awarded to Philipp Koellinger). This study makes use of data from dbGaP (accessions: phs000090 and phs000091), UK10K project (EGA accessions: EGAS00001000108 and EGAS00001000090), and UK Biobank Resource (application number: 12514). A full list of acknowledgements for these datasets can be found in part 19 of the Supplementary Note.

Author contributions

J.Y., P.M.V., and R.d.V. conceived the study. J.Y., J.Z., and P.M.V. designed the experiment. J.Z. derived the analytical methods, conducted all analyses, and developed the software with assistance and guidance from J.Y., Y.W., M.R.R., L.R.L.-J., L.Y., C.X.Y., A.X., and J.S. L.R.L.-J., A.F.M., J.E.P., G.W.M., A.M., T.E., G.G., N.R.W., and P.M.V. provided the CAGE data. J.Z. and J.Y. wrote the manuscript with the participation of all authors. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0101-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

The BayesS model. BayesS is a Bayesian MLM that simultaneously fits all the SNP effects as random:

$$\boldsymbol{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of phenotypes, μ is the fixed effect, $\mathbf{X}$ is the matrix of SNP genotype scores centered by the column means, $\boldsymbol{\beta}$ is the vector of SNP effects, and $\boldsymbol{e}$ is the residuals. The fixed effect has a flat prior: $\mu \propto constant$. In practice, we fitted principal components and other covariates in the model as fixed effects. It is common to standardize the SNP genotypes such that each column of $\mathbf{X}$ has variance 1. In BayesS, however, the SNP genotypes are not standardized, as the standardization implicitly assumes a strong negative relationship between effect size and MAF ($S=-1$)[21–23,57]. We assume that the SNP effect β_j has a hierarchical mixture prior (see Method overview):

$$\beta_j \sim N\left(0, [2p_j(1-p_j)]^S\sigma_\beta^2\right)\pi + \phi(1-\pi)$$

where ϕ is a point mass at zero and π, the proportion of SNPs with nonzero effects, reflects the polygenicity of a trait. We allow data to determine the polygenicity by assuming a uniform prior:

$$\pi \sim U(0,1)$$

The variance of SNP effects, which quantifies our prior belief on the effect size, is modeled to be related to MAF p_j through S, which is assumed to have a normal prior:

$$S \sim N(0, \sigma_S^2)$$

Namely, we a priori believe a neutral model with some certainty (quantified by the given variance) to allow the detection of selection from the data. We used $\sigma_S^2 = 1$ as the prior in the analysis of UKB traits, but used a more informative prior $\sigma_S^2 = 0.1$ in the analysis of gene expression traits in CAGE to shrink the noise more heavily toward zero given the much smaller sample size of CAGE compared with that of UKB. The prior for the common variance factor is

$$\sigma_\beta^2 \sim \nu_\beta\tau_\beta^2\chi_{\nu_\beta}^{-2}$$

where $\nu_\beta = 4$ and τ_β^2 is computed utilizing the characteristic of the distribution: if $\sigma^2 \sim \nu\tau^2\chi_\nu^{-2}$, then $\mathrm{E}(\sigma^2) = \nu\tau^2/(\nu-2)$. Rearranging the equation gives

$$\tau_\beta^2 = \frac{\nu_\beta - 2}{\nu_\beta}E(\sigma_\beta^2)$$

where

$$E(\sigma_\beta^2) = \frac{V_g}{\pi_0\sum_j[2p_j(1-p_j)]^{1+S_0}}$$

with V_g, π_0 and S_0 being the priors of the genetic variance, π and S, respectively. To remove the dependency of the hyper-parameter τ_β^2 on the prior values of the genetic variance, π and S, we compute τ_β^2 deterministically using the sampled values of these parameters for the first 2,000 MCMC cycles, and then set τ_β^2 equal to the average value across these cycles. Likewise, the prior for the residual variance is

$$\sigma_e^2 \sim \nu_e\tau_e^2\chi_{\nu_e}^{-2}$$

where $\nu_e = 4$ and $\tau_e^2 = \frac{\nu_e - 2}{\nu_e}V_e$ with V_e being a prior of the residual variance. Note that when $S=0$, our model becomes BayesCπ[14], a method that has been widely used for genomic prediction in agriculture, or Bayesian variable selection regression (BVSR) in statistics literature[58]. The sampling scheme of the parameters is given in part 2 of the Supplementary Note. We used posterior mode, standard deviation (s.d.) or highest probability density (HPD) computed from the MCMC samples to estimate the parameter (θ), standard error (s.e.) or credible interval, respectively.

The nested BayesS model is developed based on a previously published method, BayesN[59], to speed up computation when a large number of SNPs are included in the analysis. In the nested BayesS, the genome is partitioned into W-kb non-overlapping segments. Each window a priori has k SNPs with nonzero effects, where W and k are some given numbers. SNPs in the same window are individually modeled as in BayesS, as well as collectively considered as a window effect with a normal-zero mixture prior. Remarkable speedups are obtained by skipping over the windows with zero effect, focusing solely on the windows that harbor genetic signals. Thus, the reduction in computing time is inversely related to the polygenicity, which is defined here as the proportion of segments with nonzero effects. When the causal variants are not observed, choosing $k > 1$ may lead to better performance in parameter estimation than BayesS, as it refines the signal of causal variant by allowing the flanking SNPs to jointly capture its effect. In this study, the length of window was set at 200 kb with 2 SNPs a priori fitted in the model. See parts 13–15 of the Supplementary Note for details on the nested BayesS and the comparison with the standard BayesS and see part 18 of the Supplementary Note for a discussion on computational efficiency.

Estimation of SNP-based heritability. We estimate the SNP-based heritability using the sampled values of SNP effects in MCMC[60]. By definition, the genetic variance is the variance of the genetic values across individuals. In each MCMC cycle, we estimate the genetic value for each individual ($\tilde{g}_i$) using SNPs with sampled nonzero effects ($\tilde{\beta}_j$):

$$\tilde{g}_i = \sum_j X_{ij}\tilde{\beta}_j$$

Then, the estimated genetic variance in the current cycle is

$$\tilde{\sigma}_g^2 = \frac{\sum_i \tilde{g}_i^2}{N} - \left(\frac{\sum_i \tilde{g}_i}{N}\right)^2$$

Conditional on the sampled value of residual variance ($\tilde{\sigma}_e^2$), the estimated SNP-based heritability is

$$\tilde{h}_{SNP}^2 = \frac{\tilde{\sigma}_g^2}{\tilde{\sigma}_g^2 + \tilde{\sigma}_e^2}$$

In this study, we used the mode of all cycles after burn-in as the estimate of SNP-based heritability.

Hypothesis test against $S=0$. The first approach is to control the family-wise type I error rate (FWER) using an asymptotic theory that the posterior distribution has a normal approximation centered at the true parameter value under certain assumptions[61] (part 4 of the Supplementary Note). The asymptotic property of the posterior distribution for S was justified by simulation with the UKB cohort, where the type I error rate was well controlled (Supplementary Fig. 1). The second approach is to control the proportion of false positives[62] (PFP) among rejections (also known as the marginal false discovery rate or mFDR[63]) based on the posterior probability given the data ($\mathcal{D}$), e.g., $\Pr(S<0\ |\mathcal{D})$. We showed by simulation that given a large sample size, rejecting $S=0$ with $\Pr(S<0\ |\mathcal{D})$ or $\Pr(S>0\ |\mathcal{D}) \geq \gamma$ guarantees PFP to be less than $1-\gamma$ (Supplementary Fig. 2). The former approach is more stringent but the advantage of the latter approach is that power is not inversely related to the number of traits tested[62].

Simulation based on genotyped array SNPs. The simulation based on ARIC + GENEVA[12] was used for testing the methods. We used PLINK 1.9[64] to carry out standard quality control (QC) procedures on the dataset, including removal of SNPs with missingness > 5%, Hardy–Weinberg equilibrium test at $P < 10^{-6}$, or MAF < 1%, and removal of individuals with missing genotypes < 1% and genetic relationship < 0.05 estimated from all SNPs after QC using GCTA–GRM[56]. After QC, a total of 12,942 unrelated individuals and 564,959 SNPs remained. A quantitative trait was simulated by choosing 1,000 SNPs at random as causal variants with their effects sampled from $N(0,1)$. To simulate a spectrum of relationships between MAF and effect size, the SNP effect was multiplied by $[2p_j(1-p_j)]^{S/2}$ where $S = -1, -0.5, 0, 0.5$, or 1, representing the negative-to-positive relationship between effect size and MAF, including the case of independence when $S=0$. An individual phenotype specific to a given value of S was generated by adding a random normal residual with variance identical to the genetic variance, giving each simulated trait a heritability of 0.5. The simulation process was repeated 100 times. We analyzed the simulated data with and without the causal variants in the model. To evaluate the robustness of our method to the starting values of parameters, we used an arbitrary value of 0 for S, 0.2 for heritability, and 0.05 for π, respectively, to start the MCMC.

Analysis of the UK Biobank data. We have access to 46 complex traits in the UK Biobank[19], in which phenotype data were collected from over 500,000 individuals aged between 40 and 69 across the United Kingdom. The interim release contains genotypes for 152,736 samples at 806,466 SNPs on a customized Affymetrix Axiom array after QC procedures[65]. We selected a subset of 140,408 individuals that had a self-reported gender identical to the genetically inferred gender and a European ethnicity derived from a principal component analysis together with self-reported ethnicity. Furthermore, we removed individuals with genomic relatedness > 0.05 estimated from all SNPs using GCTA–GRM[56] and SNPs with genotype missing rate > 5%, Hardy–Weinberg equilibrium test $P < 10^{-6}$, or MAF < 1%. The final dataset consisted of 126,752 individuals of European ancestry with 483,634 common SNPs (MAF > 1%). After removal of five duplicated traits and five traits with sample size (N) < 20,000, 36 traits remained for analysis, including 32 quantitative traits (anthropometric, cardiovascular and reproductive),

2 categorical traits (male pattern baldness (MPB) and years of schooling (educational attainment)) and 2 diseases or disorders (type 2 diabetes (T2D) and depression (DEP)). The sample size for each trait is shown in Supplementary Table 2; most traits had $N > 100{,}000$. The prevalence of T2D and DEP in the sample was 5.35% and 6.70%, respectively. For T2D and DEP, the estimates of SNP-based heritability were on the liability scale and were converted from the observed scale[10], assuming population prevalences of 8%[35] and 15%[66], respectively. The phenotypes of quantitative traits were standardized within each sex group after regressing out the age effect. For educational attainment, the years of schooling are pre-adjusted by sex, a third order polynomial of year-of-birth, and year-of-birth by sex interactions. We used BayesS for the analysis, where the first 20 principal components (PC) of GRM were fitted as fixed effects to account for the effects due to population stratification. For the disease traits, sex and age were fitted as covariates in addition to PCs, and for MPB, only age was fitted as the additional covariate. The diagnostic results of MCMC convergence in BayesS are shown in Supplementary Fig. 26 and Supplementary Table 9.

Simulation study based on UK10K sequence data. We included 475,314 sequence variants (51.4% are rare, MAF < 1%) from chromosomes 21 and 22 of the whole-genome sequencing (WGS) data from the UK10K project[26]. We randomly sampled 100 causal variants from all sequence variants or the variants at DNase I-hypersensitive sites (DHSs). We also simulated a more polygenic architecture with 1,000 causal variants. Furthermore, the sequence variants were downsampled to match the SNPs on the Affymetrix Axiom chip (14,631 SNPs with MAF > 1%). The effects of causal variants were generated given a spectrum of S values as in the ARIC + GENEVA simulation (Methods). A heritability of 0.8 was used to compensate for the small sample size ($N = 3{,}642$).

Coalescent simulation of sequence variants with a large sample size. We used the coalescent simulation software msprime[41] to efficiently generate genotypes of sequence variants on a 10-Mb segment for 100,000 individuals. Under the demographic model described in Gravel et al.[47], we observed an excess of rare variants, as expected (Supplementary Fig. 7b). We randomly chose 100 sequence variants as the causal variants, the effects of which were sampled given a spectrum of S values as in the ARIC + GENEVA simulation (Methods). Three levels of heritability (0.01, 0.05, and 0.1) were investigated in the simulation.

Forward simulation for different types of natural selection. We ran forward simulations using SLiM[67] to confirm that the relationship between effect size and MAF is subject to different types of natural selection. We simulated a 10-Mb region where new mutations occurred with probability of 0.95 to be neutral and of 0.05 to be a causal variant with an effect sampled from $N(0,1)$. The mutation rate was set at 1.65×10^{-8} (ref. [68]). The phenotype of an individual was simulated based on the genotypic values with a heritability of 0.1 at all segregating causal variants in the current generation. We simulated the evolution of a population of 1,000 individuals over 10,000 generations (this is equivalent to 100,000 generations in a population of 10,000 individuals[69]). The first 5,000 generations were used as a burn-in period, in which the phenotype did not affect fitness and all variants (including the causal variants) were under neutral variation. From generation 5,001, we related the standardized phenotype, with mean 0 and variance 1, to fitness through hypothetical functions that represent different types of selection (Fig. 5). For directional selection, the phenotype was positively or negatively correlated to fitness through a simple linear function. For stabilizing selection, we used a normal curve to model fitness achieving an optimum at at intermediate phenotype value. For disruptive selection, a reversed normal curve was used to model the phenotypes at the two tails producing the highest fitness. In the last generation of selection, we investigated the joint distribution of effects and frequencies of the derived alleles, the joint distribution of effects and frequencies of the coded alleles (arbitrarily chosen as in reality where derived alleles are unknown), and the relationship between the variance of the coded-allele effects and MAF. We collected results from 200 simulation replicates.

Consortium for the Architecture of Gene Expression (CAGE) dataset. We analyzed the mRNA levels for 36,778 transcript expression traits (probes) from the CAGE[20] dataset using the nested BayesS method. The CAGE data comprised measurements from 36,778 gene expression probes in peripheral blood, with a subset of 1,748 unrelated (genomic relatedness < 0.05) European individuals from the total 2,765 individuals used for this analysis. Full details of the cohorts contributing to CAGE, and their sample preparation, normalization and genotype imputation are outlined in Lloyd-Jones et al.[20]. Briefly, the quantification of gene expression for each cohort was performed using the Illumina Whole-Genome Expression BeadChips. The gene expression levels in each cohort were initially normalized, followed by a quantile adjustment to standardize the distribution of expression levels across samples. We corrected for age, gender, cell counts, and batch effects as well as hidden heterogeneous sources of variability. The rank-based inverse-normal transformation was used to normalize the measurements for each probe to be normally distributed with mean 0 and variance 1. Probes measuring expression levels of genes located on chromosomes X and Y were removed from the analysis. The initial CAGE dataset consisted of seven unique cohorts that were genotyped on different SNP arrays. Therefore, genotype data were imputed to the 1,000 Genomes Phase 1 reference panel[70], resulting in 7,763,174 SNPs passing quality control, of which 1,066,738 SNPs overlapped with HapMap3 and were used for analysis.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. This study makes use of data from dbGaP (accessions: phs000090 and phs000091), UK10K project (EGA accessions: EGAS00001000108 and EGAS00001000090), and UK Biobank Resource (application number: 12514). As per the ethics agreement of the CAGE consortium, all raw and normalized genotype and expression data are available to consortium members. Consortium membership is open, but requires approval from the steering committee.

Code availability. BayesS has been implemented in a software tool called GCTB (genome-wide complex trait Bayesian analyses), the computer code for which is freely available at http://cnsgenomics.com/software/gctb.

References

57. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
58. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815 (2011).
59. Zeng, J. Whole Genome Analyses Accounting for Structures in Genotype Data. PhD dissertation, Iowa State University, Chapter 2, 6-33 (2015).
60. Fernando, R. L. & Garrick, D. Bayesian methods applied to GWAS. *Methods Mol. Biol.* **1019**, 237–274 (2013).
61. Gelman, A. et al. *Bayesian Data Analysis*. (CRC Press, Boca Raton, FL, 2014).
62. Fernando, R. L. et al. Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**, 611–619 (2004).
63. Storey, J. D. The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **69**, 347–368 (2007).
64. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
65. UK Biobank Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource. UK Biobank http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf (2015).
66. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).
67. Messer, P. W. SLiM: simulating evolution with selection and linkage. *Genetics* **194**, 1037–1039 (2013).
68. Palamara, P. F. et al. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am. J. Hum. Genet.* **97**, 775–789 (2015).
69. Enard, D., Messer, P. W. & Petrov, D. A. Genome-wide signals of positive selection in human evolution. *Genome Res.* **24**, 885–895 (2014).
70. Abecasis, G. R. et al. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

nature research

Corresponding Author: Jian Yang

Date: Feb 23, 2018

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   We applied our method to complex traits and gene expression traits. For the complex trait analyses, we chose traits with sample size greater than 20,000, and most of the traits have sample size greater than 100,000. For the analyses of gene expression traits, the total sample size was 1,748. We have quantified the statistical power of our analysis by simulations.

2. Data exclusions

   Describe any data exclusions.

   Our study was restricted to data sets of genetically unrelated individuals of European ancestry (genomic relationship < 0.05) and common SNPs (minor allele frequency > 1%). Traits for which there was evidence for lack of convergence for MCMC were excluded (based on the Geweke test and Heidelberger and Welch's convergence diagnostic test).

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   The experimental replication was not attempted.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   This is not relevant to our study because we do not have treatment/ control groups.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   This is not relevant to our study because we do not have treatment/ control groups.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	We implemented our method in a software GCTB. The code is freely available at http://cnsgenomics.com/software/gctab. We also used PLINK v1.9 for data quality control, GCTA v1.26.0 for the GREML-LDMS analysis, LDSC v1.0.0 for the LD score regression analysis, SLiM v2.3 for the forward simulation, and msprime v0.5.0 for the coalescence simulation. We have provided the links to these software tools in the URLs section.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	The use of UK Biobank data was in concordance with the UK Biobank policy. The genotype and expression data in CAGE data set are available to consortium members. Consortium membership is open, but requires approval from the steering committee.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The human research participants in the UK Biobank have age between 40 and 69 across the UK. The genotypic information are obtained with a customized Affymetrix Axiom array. We selected a subset of individuals with European ethnicity and negligible genomic relatedness for the analysis. The CAGE samples consisted of data from five cohorts, where the genotype data were obtained using different genotyping platforms and were imputed to the 1000 Genomes Phase 1 Version 3 reference panel. The same selection criteria on the ethnicity and genomic relationship as in UK Biobank samples were applied to the CAGE samples.